Multiple Regression

# A. Business Question

### What is the relationship between yearly customer data usage and other customer demographic variables?

To better position infrastructure and resources based on data demand, a telecommunications organization would benefit from a predictive modeling analysis estimating yearly customer data usage based on various customer variables.

# B. Multiple Regression Model

Using a multiple regression model we can explore the relationship between customer variables and customer data usage. The multiple regression model is chosen as we have multiple variables for which to explore the relationships. A multiple regression model will predict a response variable’s value given a number of explanatory variables. As in this case the goal is to predict the extent of yearly data usage demand, the multiple regression model is the appropriate technique.

Using a multiple regression model we have an assumption that the explanatory variables are not highly correlated with each other. Also, that there is a linear relationship between the response variable and the explanatory variables. Further, we assume that the residuals, being the difference between the predicted values and the observed values of the multiple regression model, are normally distributed.

R will be used for this analysis. R is open source software that was specifically made for statistical analysis. Using R, we can ingest the raw data set, and leveraging an extensive library of data manipulation and visualization packages, clean and investigate the data. More information can be found on the R project website (<https://www.r-project.org/>).

# C. Data Preparation

To prepare the data, first a check is run for missing values in the data set. No action action is needed as the data set does not contain missing values.

df<-read.csv("c:/users/shua/documents/Predictive Modeling\_D208/churn\_clean.csv")  
sapply(df, function(x) sum(is.na(x)))

## CaseOrder Customer\_id Interaction   
## 0 0 0   
## UID City State   
## 0 0 0   
## County Zip Lat   
## 0 0 0   
## Lng Population Area   
## 0 0 0   
## TimeZone Job Children   
## 0 0 0   
## Age Income Marital   
## 0 0 0   
## Gender Churn Outage\_sec\_perweek   
## 0 0 0   
## Email Contacts Yearly\_equip\_failure   
## 0 0 0   
## Techie Contract Port\_modem   
## 0 0 0   
## Tablet InternetService Phone   
## 0 0 0   
## Multiple OnlineSecurity OnlineBackup   
## 0 0 0   
## DeviceProtection TechSupport StreamingTV   
## 0 0 0   
## StreamingMovies PaperlessBilling PaymentMethod   
## 0 0 0   
## Tenure MonthlyCharge Bandwidth\_GB\_Year   
## 0 0 0   
## Item1 Item2 Item3   
## 0 0 0   
## Item4 Item5 Item6   
## 0 0 0   
## Item7 Item8   
## 0 0

In order to streamline the data set only select variables will be used. Starting broadly, the data set will be filtered to only include the following variables: Population, Children, Age, Outage seconds per week, Contacts, Multiple, Streaming TV, Streaming Movies, Tenure, Monthly Charge, and Bandwidth GB per Year.

colnames(df)

## [1] "CaseOrder" "Customer\_id" "Interaction"   
## [4] "UID" "City" "State"   
## [7] "County" "Zip" "Lat"   
## [10] "Lng" "Population" "Area"   
## [13] "TimeZone" "Job" "Children"   
## [16] "Age" "Income" "Marital"   
## [19] "Gender" "Churn" "Outage\_sec\_perweek"   
## [22] "Email" "Contacts" "Yearly\_equip\_failure"  
## [25] "Techie" "Contract" "Port\_modem"   
## [28] "Tablet" "InternetService" "Phone"   
## [31] "Multiple" "OnlineSecurity" "OnlineBackup"   
## [34] "DeviceProtection" "TechSupport" "StreamingTV"   
## [37] "StreamingMovies" "PaperlessBilling" "PaymentMethod"   
## [40] "Tenure" "MonthlyCharge" "Bandwidth\_GB\_Year"   
## [43] "Item1" "Item2" "Item3"   
## [46] "Item4" "Item5" "Item6"   
## [49] "Item7" "Item8"

df<-df[,c(11,15,16,21,23,31,36,37,40,41,42)]  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Outage\_sec\_perweek" "Contacts" "Multiple"   
## [7] "StreamingTV" "StreamingMovies" "Tenure"   
## [10] "MonthlyCharge" "Bandwidth\_GB\_Year"

Next, we will investigate the data types for each of the selected variables to understand which are categorical and which are continuous or discrete.

The following three variables are found to be categorical and will need to be transformed to dummy numeric variables: Multiple, StreamingTV, and StreamingMovies

To do this we will use the case\_when statement from the “dplyr” library. Dplyr is a data manipulation package within R. More information can be found on the tidyverse website (<https://dplyr.tidyverse.org/>).

str(df)

## 'data.frame': 10000 obs. of 11 variables:  
## $ Population : int 38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...  
## $ Children : int 0 1 4 1 0 3 0 2 2 1 ...  
## $ Age : int 68 27 50 48 83 83 79 30 49 86 ...  
## $ Outage\_sec\_perweek: num 7.98 11.7 10.75 14.91 8.15 ...  
## $ Contacts : int 0 0 0 2 2 3 0 0 2 1 ...  
## $ Multiple : chr "No" "Yes" "Yes" "No" ...  
## $ StreamingTV : chr "No" "Yes" "No" "Yes" ...  
## $ StreamingMovies : chr "Yes" "Yes" "Yes" "No" ...  
## $ Tenure : num 6.8 1.16 15.75 17.09 1.67 ...  
## $ MonthlyCharge : num 172 243 160 120 150 ...  
## $ Bandwidth\_GB\_Year : num 905 801 2055 2165 271 ...

library(dplyr)

## Warning: replacing previous import 'vctrs::data\_frame' by 'tibble::data\_frame'  
## when loading 'dplyr'

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

df$dummy\_multiple<-case\_when(  
 df$Multiple == "Yes" ~ 1,  
 TRUE ~ 0  
)  
df$dummy\_streamingTV<-case\_when(  
 df$StreamingTV == "Yes" ~ 1,  
 TRUE ~ 0  
)  
df$dummy\_streamingMovies<-case\_when(  
 df$StreamingMovies == "Yes" ~ 1,  
 TRUE ~ 0  
)

Next, summary statistics will be generated for each of the numeric variables to gain an understanding of variation and central tendency. No further action is needed in this regard as none of the values appear unreasonable.

# Summary statistics for each variable  
colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Outage\_sec\_perweek" "Contacts" "Multiple"   
## [7] "StreamingTV" "StreamingMovies" "Tenure"   
## [10] "MonthlyCharge" "Bandwidth\_GB\_Year" "dummy\_multiple"   
## [13] "dummy\_streamingTV" "dummy\_streamingMovies"

summary(df$Population)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 738 2910 9757 13168 111850

summary(df$Children)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 1.000 2.088 3.000 10.000

summary(df$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 35.00 53.00 53.08 71.00 89.00

summary(df$Outage\_sec\_perweek)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.09975 8.01821 10.01856 10.00185 11.96949 21.20723

summary(df$Contacts)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 1.0000 0.9942 2.0000 7.0000

summary(df$Tenure)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 7.918 35.431 34.526 61.480 71.999

summary(df$MonthlyCharge)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 79.98 139.98 167.48 172.62 200.73 290.16

summary(df$Bandwidth\_GB\_Year)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 155.5 1236.5 3279.5 3392.3 5586.1 7159.0

Next, visualizations are produced to check the data preparedness and inspect for anomalies. First, univariate visualizations of select variables do not show anything that require attention.

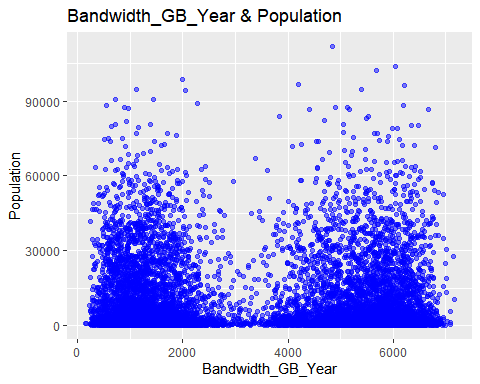
Finally, bivariate visualizations between the independent variables and the defendant variable (bandwidth\_gb\_year) are produced . For this task the library “ggplot2” is used which is useful for plotting data within R. More information on the library ggplot2 can be found on the tidyverse website (<https://ggplot2.tidyverse.org/>). After review of the bivariate visualizations the data preparedness is determined to be sufficient.

Interestingly, a strong correlation is suggested between Tenure and Bandwidth GB per year.

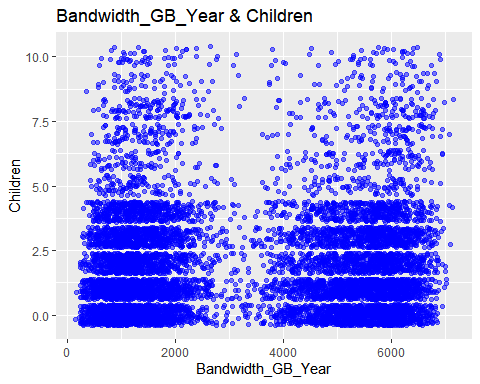
# Scatter plots for each variable  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

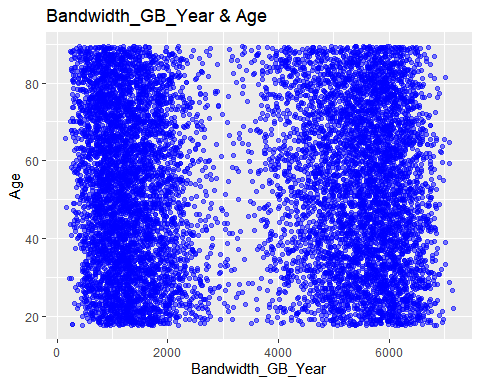
ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Population))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Population")



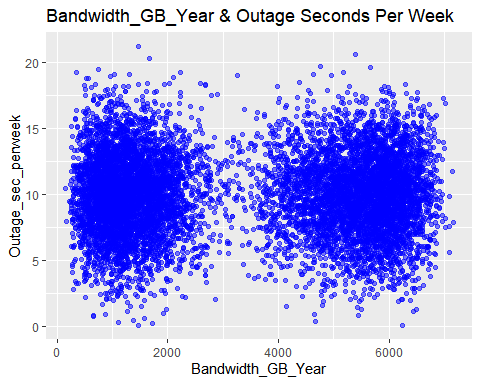
ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Children))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Children")



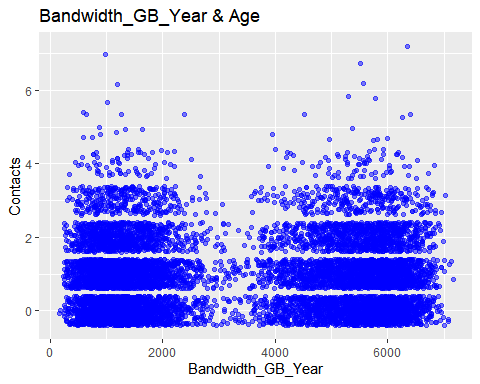
ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Age))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Age")



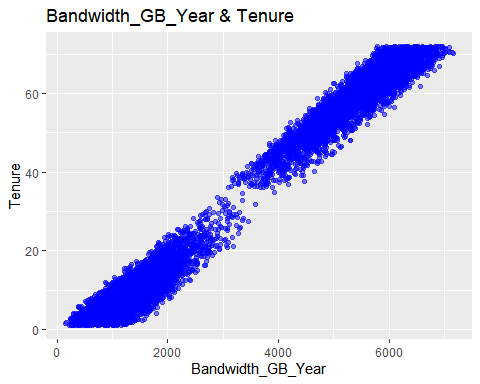
ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Outage\_sec\_perweek))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Outage Seconds Per Week")



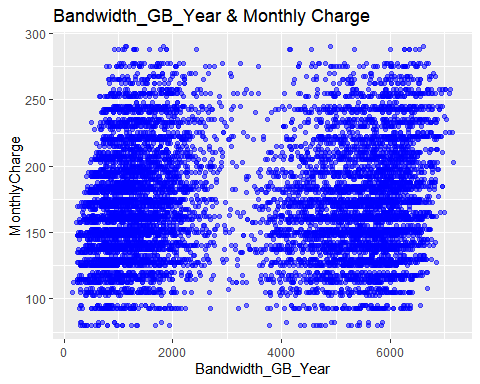
ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Contacts))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Age")



ggplot(df, aes(x=Bandwidth\_GB\_Year, y=Tenure))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Tenure")



ggplot(df, aes(x=Bandwidth\_GB\_Year, y=MonthlyCharge))+  
 geom\_point(position="jitter", color="blue", alpha=.5)+  
 ggtitle("Bandwidth\_GB\_Year & Monthly Charge")



The following prepared data set will be used for the analysis going forward: “prepared\_churn\_clean.csv”.

write.csv(df, "c:/users/shua/documents/Predictive Modeling\_D208/prepared\_churn\_clean.csv")

# D. Model Comparison and analysis

An initial multiple regression model is prepared using all of the variables identified during the data preparation phase. The result is a model having a multiple R Squared value of .9904. This value means that 99% of variation in the data set is explained by the initial multiple regression model. That is fantastic!

However, in our bivariate visualizations during the data preparation phase, we noted that Tenure demonstrated a strong visual correlative relationship with the dependent variable. As such, we can likely reduce this model while preserving a high R Squared value. Looking at the p values in the coefficients section of the model summary, we see the following variables annotated as statistically significant:

1. Children
2. Age
3. Multiple
4. StreamingTV
5. StreamingMovies
6. Tenure
7. Monthly Charge

We will therefore attempt to reduce the model by dropping the Outage seconds per week, population, and contacts variables which did not demonstrate a statistically significant p value. This leaves us with only the 7 independent/explanatory variables listed above.

colnames(df)

## [1] "Population" "Children" "Age"   
## [4] "Outage\_sec\_perweek" "Contacts" "Multiple"   
## [7] "StreamingTV" "StreamingMovies" "Tenure"   
## [10] "MonthlyCharge" "Bandwidth\_GB\_Year" "dummy\_multiple"   
## [13] "dummy\_streamingTV" "dummy\_streamingMovies"

initial\_lm<-lm(Bandwidth\_GB\_Year ~ Population + Children + Age + Outage\_sec\_perweek + Contacts + dummy\_multiple + dummy\_streamingTV + dummy\_streamingMovies + Tenure + MonthlyCharge, data=df)  
summary(initial\_lm)

##   
## Call:  
## lm(formula = Bandwidth\_GB\_Year ~ Population + Children + Age +   
## Outage\_sec\_perweek + Contacts + dummy\_multiple + dummy\_streamingTV +   
## dummy\_streamingMovies + Tenure + MonthlyCharge, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -328.44 -170.29 -86.38 220.14 483.30   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.589e+02 1.552e+01 23.132 < 2e-16 \*\*\*  
## Population 1.157e-04 1.486e-04 0.778 0.436   
## Children 3.087e+01 9.998e-01 30.872 < 2e-16 \*\*\*  
## Age -3.291e+00 1.037e-01 -31.735 < 2e-16 \*\*\*  
## Outage\_sec\_perweek -4.659e-01 7.210e-01 -0.646 0.518   
## Contacts 2.839e+00 2.171e+00 1.307 0.191   
## dummy\_multiple 5.811e+01 5.432e+00 10.698 < 2e-16 \*\*\*  
## dummy\_streamingTV 2.071e+02 6.018e+00 34.410 < 2e-16 \*\*\*  
## dummy\_streamingMovies 1.850e+02 6.814e+00 27.153 < 2e-16 \*\*\*  
## Tenure 8.201e+01 8.113e-02 1010.864 < 2e-16 \*\*\*  
## MonthlyCharge 5.422e-01 1.008e-01 5.381 7.57e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 214.5 on 9989 degrees of freedom  
## Multiple R-squared: 0.9904, Adjusted R-squared: 0.9904   
## F-statistic: 1.028e+05 on 10 and 9989 DF, p-value: < 2.2e-16

The results of the reduced model retained a 0.9904 multiple R squared value. This indicates that the reduced model still explains 99% of the variation in the data set while relying on fewer independent variables than the initial regression model.

library(ggfortify)

## Warning: package 'ggfortify' was built under R version 4.0.5

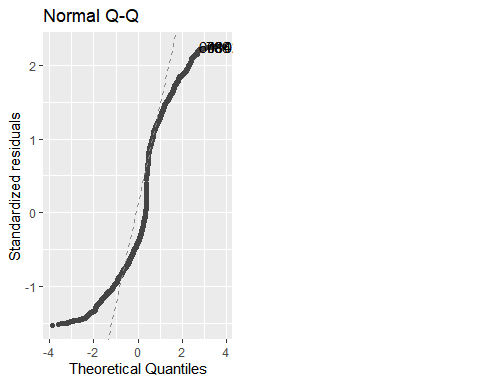
reduced\_lm<-lm(Bandwidth\_GB\_Year ~ Children + Age + dummy\_multiple + MonthlyCharge + dummy\_streamingTV + dummy\_streamingMovies + Tenure, data=df)  
summary(initial\_lm)

##   
## Call:  
## lm(formula = Bandwidth\_GB\_Year ~ Population + Children + Age +   
## Outage\_sec\_perweek + Contacts + dummy\_multiple + dummy\_streamingTV +   
## dummy\_streamingMovies + Tenure + MonthlyCharge, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -328.44 -170.29 -86.38 220.14 483.30   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.589e+02 1.552e+01 23.132 < 2e-16 \*\*\*  
## Population 1.157e-04 1.486e-04 0.778 0.436   
## Children 3.087e+01 9.998e-01 30.872 < 2e-16 \*\*\*  
## Age -3.291e+00 1.037e-01 -31.735 < 2e-16 \*\*\*  
## Outage\_sec\_perweek -4.659e-01 7.210e-01 -0.646 0.518   
## Contacts 2.839e+00 2.171e+00 1.307 0.191   
## dummy\_multiple 5.811e+01 5.432e+00 10.698 < 2e-16 \*\*\*  
## dummy\_streamingTV 2.071e+02 6.018e+00 34.410 < 2e-16 \*\*\*  
## dummy\_streamingMovies 1.850e+02 6.814e+00 27.153 < 2e-16 \*\*\*  
## Tenure 8.201e+01 8.113e-02 1010.864 < 2e-16 \*\*\*  
## MonthlyCharge 5.422e-01 1.008e-01 5.381 7.57e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 214.5 on 9989 degrees of freedom  
## Multiple R-squared: 0.9904, Adjusted R-squared: 0.9904   
## F-statistic: 1.028e+05 on 10 and 9989 DF, p-value: < 2.2e-16

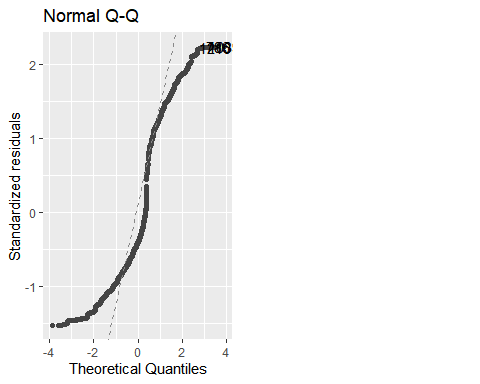
summary(reduced\_lm)

##   
## Call:  
## lm(formula = Bandwidth\_GB\_Year ~ Children + Age + dummy\_multiple +   
## MonthlyCharge + dummy\_streamingTV + dummy\_streamingMovies +   
## Tenure, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -327.92 -170.33 -86.25 218.91 480.93   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 358.07127 13.56753 26.392 < 2e-16 \*\*\*  
## Children 30.83359 0.99958 30.847 < 2e-16 \*\*\*  
## Age -3.28804 0.10369 -31.710 < 2e-16 \*\*\*  
## dummy\_multiple 57.94389 5.43123 10.669 < 2e-16 \*\*\*  
## MonthlyCharge 0.54344 0.10076 5.393 7.08e-08 \*\*\*  
## dummy\_streamingTV 206.98257 6.01744 34.397 < 2e-16 \*\*\*  
## dummy\_streamingMovies 184.94009 6.81298 27.145 < 2e-16 \*\*\*  
## Tenure 82.00890 0.08113 1010.892 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 214.5 on 9992 degrees of freedom  
## Multiple R-squared: 0.9904, Adjusted R-squared: 0.9904   
## F-statistic: 1.469e+05 on 7 and 9992 DF, p-value: < 2.2e-16

ggplot2::autoplot(initial\_lm, which=2)



ggplot2::autoplot(reduced\_lm, which=2)



# E. Analysis

The initial regression model had a wide selection of variables that might help predict customer yearly data usage. Variables along with rationale for selection are as follows:

1. Population - customers in more heavily populated areas would have higher quality service and therefore tend to use more data
2. Children - data usage increases as the number of children increases
3. Age - data usage increases as age decreases
4. Outage seconds per week - data outages might increase with higher data usage
5. Contacts - a customer with higher data usage might contact customer support a greater number of times
6. Multiple - a customer with multiple lines would have higher data usage
7. Streaming TV - streaming services would lead to higher data usage
8. Streaming Movies - streaming services would lead to higher data usage
9. Tenure - customer data usage might be higher with less tenured customers due to special marketing offers
10. Monthly Charge - higher monthly charge would indicate need for services with higher data usage
11. Bandwidth GB per Year - the dependent variable

It was observed during the bivariate visualization analysis that customer tenure had a strong linear correlative relationship with data usage. The initial model confirmed this observation, along with 6 other variables also having a statistically significant p value. These 7 statistically significant variables are used in the reduced model.

The reduced model maintains a 99% multiple R squared value while relying on fewer predictor variables.

The model fit can be visualized by plotting residuals on a Q-Q plot. The “ggfortify” package will be used which is a useful extension to the ggplot2 package. More information on the ggfortify package can be found on the Cran R project website (“<https://cran.r-project.org/web/packages/ggfortify/index.html>”)

In this case, the points on the QQ plot do not track along the straight line. This indicates that the model residuals do not have a normal distribution.

# F. Summary

The results of the reduced model show that tenure is a very good predictor for yearly customer data usage.

The model coefficients show that for every 1 child the data usage GB per year would increase by 31. For every 1 year of age the data usage would decrease by 3 GB. For customer having multiple lines the data usage would increase by 58 GB. For every 1 dollar in monthly charge the data usage would increase by half a GB. For customers having streaming TV the data usage GB per year would increase by 206. For customers having streaming movies the data usage GB per year would increase by 184. Finally, for every 1 month that the customer has stayed with the provider the data usage GB per year would increase by 82.

The regression equation for the reduced model therefore is Bandwidth\_B\_Year = intercept of 358 + Children*30.83 - Age*3.28 + dummy\_multiple*57.94 + monthlyCharge*.54 + dummyStreamingTV*206.98 + dummy\_streamingMovies*184.94 + Tenure\*82.

The model fit visualization however indicates that the model residuals do not follow a normal distribution. This of course violates our beginning assumption that the residuals of a multiple regression model would follow a normal distribution.

The model is limited to the data set provided. It would be both interesting and beneficial to understand additional factors such as the ages of the children as well as whether the customer joined the provider on any type of offertory rate. Perhaps these variables would prove statistically significant and result in a better fitted model.

The model will be useful to the organization by helping predict yearly customer data usage. It is recommended to use this model in planning infrastructure maintenance and improvements accordingly to meet the data and servicing demands of their customer base, granted that different variables can be identified for a better fitted model.